Briefings in Bioinformatics Advance Access published March 12, 2016

Briefings in Bioinformatics, 2016, 1–9

doi: 10.1093/bib/bbw021
Paper

Prediction and identification of the effectors of heterotrimeric G proteins in rice (*Oryza sativa* L.)

Kuan Li,* Chaoqun Xu,* Jian Huang,* Wei Liu, Lina Zhang, Weifeng Wan, Huan Tao, Ling Li, Shoukai Lin, Andrew Harrison and Huaqin He

Corresponding author. Huaqin He, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China. E-mail: hehq3@fafu.edu.cn
*These authors contributed equally to this work.

Abstract

Heterotrimeric G protein signaling cascades are one of the primary metazoan sensing mechanisms linking a cell to environment. However, the number of experimentally identified effectors of G protein in plant is limited. We have therefore studied which tools are best suited for predicting G protein effectors in rice. Here, we compared the predicting performance of four classifiers with eight different encoding schemes on the effectors of G proteins by using 10-fold cross-validation. Four methods were evaluated: random forest, naive Bayes, K-nearest neighbors and support vector machine. We applied these methods to experimentally identified effectors of G proteins and randomly selected non-effector proteins, and tested their sensitivity and specificity. The result showed that random forest classifier with composition of K-spaced amino acid pairs and composition of motif or domain (CKSAAP_PROSITE_200) combination method yielded the best performance, with accuracy and the Mathew's correlation coefficient reaching 74.62% and 0.49, respectively. We have developed G-Effector, an online predictor, which outperforms BLAST, PSI-BLAST and HMMER on predicting the effectors of G proteins. This provided valuable guidance for the researchers to select classifiers combined with different feature selection encoding schemes. We used G-Effector to screen the effectors of G protein in rice, and confirmed the candidate effectors by gene co-expression data. Interestingly, one of the top 15 candidates, which did not appear in the training data set, was validated in a previous research work. Therefore, the candidate effectors list in this article provides both a clue for researchers as to their function and a framework of validation for future experimental work. It is accessible at http://bioinformatics.fafu.edu.cn/geffector.

Key words: rice (*Oryza sativa* L.); heterotrimeric G proteins; effectors; predicting

Introduction

The maintenance of homeostasis in a living organism is fine-tuned by the communication between cell and environment. This helps cells to survive in unfavorable environment and under stressful conditions [1]. One of the primary sensing and physiologically important mechanisms used by metazoans is heterotrimeric G protein (G protein) signaling cascades [2]. This system is composed of a plasma membrane localized G-protein coupled receptors (GPCRs) that transfer the extracellular signal to an intracellular G protein, which in turn activate the

Kuan Li is an MSc student at the College of Life Sciences, Fujian Agriculture and Forestry University, China. His research focuses on developing a new predictor on the effectors of heterotrimeric G proteins.
Chaoqun Xu is an MSc student at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Jian Huang is a PhD student at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Wei Liu is a biologist at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Lina Zhang is an assistant professor at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Weifeng Wan is an MSc student at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Huan Tao is an MSc student at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Ling Li is an MSc student at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Shoukai Lin is a PhD student at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Andrew Harrison is a senior lecturer in bioinformatics at Department of Mathematical Sciences, University of Essex, UK.
Huaqin He is a professor of bioinformatics at the College of Life Sciences, Fujian Agriculture and Forestry University, China.
Submitted: 1 December 2015; **Received (in revised form):** 4 February 2016

downstream effectors and signaling cascades, thereby causing defense responses [1, 2].

Heterotrimeric G protein consists of three subunits, α, ß and γ (named Gα, Gß and Gγ, respectively), which form a heterotrimeric complex in the inactive state. When an agonist binds to its specific GPCR, an inactive G protein switches to its active conformation. As a result, Gα-Guanosine Triphosphate (GTP) separates from the Gßγ dimer and both Gα-GTP and the Gßγ dimer activate downstream effectors. The GTP that is bound to Gα is then hydrolyzed to Guanosine Diphosphate (GDP), thereby inactivating Gα and allowing its reassociation with the Gßγ dimer to reform the inactive heterotrimeric complex.

Many new GPCRs have been identified in metazoans in the past decades. In the gpDB database (Database of G proteins, GPCR and Effectors), there are 2738 GPCRs and 1390 effectors in 469 species [3]. Whole-genome sequencing efforts have shown that heterotrimeric G-protein signaling can be highly complex. There are 23 Gα, 5 Gß and 12 Gγ subunits in the human proteome [4], leading to over 1300 theoretical heterotrimeric complexes [2].

The number of heterotrimeric signaling complex components in plants, however, is dramatically less than that in human. There are only one canonical Gα subunit, one Gß subunit and two identified Gγ subunits in the two model plants, Arabidopsis and rice [5]. Searches of gpDB databases did not identify any plant sequences in the GPCR and effectors category [2]. For the past decade, there has been only one putative GPCR (GCR1) identified and experimentally investigated in Arabidopsis [6]. GCR2 was reported as a new GPCR in Arabidopsis [7], although it does not appear to have the canonical seven transmembrane topology of known GPCRs [8]. In rice (*Oryza sativa L.*), only a putative GPCR was isolated and functioned to promote stress tolerance [1].

Many comprehensive bioinformatics methods have been developed to predict and characterize potential GPCRs [3]. More than 850 proteins were predicted as human GPCRs [9]. Moriyama *et al.* [10] used multiple computational methods, along with HMMTOP2, to identify 54 GPCR candidates in *Arabidopsis*, whereas Gookie *et al.* [2] used a combinatorial approach to identify novel GPCRs within *Arabidopsis*, *Oryza*, and *Populus* proteomes.

Although GPCRs and their effectors are the two key components of G protein signaling cascades, the research work on the effectors of G proteins is limited when compared with the research on GPCRs. To the best of our knowledge, there are few effectors experimentally identified in plants. There are some examples, such as acireductone dioxygenase 1 that was recently found to be an effector of Gß in *Arabidopsis* [11]. Furthermore, there are no specific predictors developed for predicting effectors of G proteins in plants. The researchers have to use the traditional similarity search tools, such as BLAST, PSI-BLAST or HMMER, to predict the effectors of G proteins. In this research work, we first evaluate the performance of different classifiers combined with different encoding schemes for feature selection. We find that random forest (RF) classifier combined with CKSAAP_PROSITE_200 for feature selection yielded the best performance. Second, we develop an online predictor, G-Effector, by using RF classifier combined with CKSAAP_PROSITE_200 for feature selection. Third, we compare the predicting performance of G-Effector with traditional tools, including BLAST, PSI-BLAST and HMMER. We have also screened the candidate G protein effectors in rice made by the new predictor. One of the top 15 candidate effectors has been reported by Bhardwaj *et al.* [12]. The candidate effectors' list in this article provides both a clue for researchers as to their function and a framework of validation by future experimental work.

Methods

Preprocessing of data set

We collected 391 subunits of G proteins in 469 species from gpDB database (http://bioinformatics.biol.uoa.gr/gpDB), whereas 153 interacting proteins of these G proteins were downloaded from DIP (Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/dip/) and Intact (ftp://ftp.ebi.ac.uk/pub/databases/intact/2011-03-03/psimitab/intact.zip). Those annotated as 'reviewed' but not GPCRs, regulators or cytoskeletal proteins remained in the data set.

We found 116 candidate effectors from 9 species: *Arabidopsis thaliana*; *Bos taurus*; *Caenorhabditis elegans*; *Dictyostelium discoideum*; *Drosophila melanogaster*; *Homo sapiens*; *Mus musculus*; *Rattus norvegicus*; and *Saccharomyces cerevisiae*. All the protein sequences in these 9 species, excluding the 116 candidate effectors, were named non-effectors and downloaded from UniProt. After filtering by CD-HIT at 40% sequence identity, 104 candidate effectors and 30,622 non-effectors protein sequences were compiled into positive and negative data sets, which could be downloaded from http://bioinformatics.fafu.edu.cn/G_effector_dataset/.

To balance the positive and negative data set during 10-fold cross-validation processes, we partitioned the negative data set into 10-folds, and randomly selected 104 sequences from each fold [13]. Subsequently, each fold of data was in turn used as the test data and the remaining 9-folds of data as the training data and so each datum was tested exactly once. The jackknife test was also used to examine the prediction performance.

Encoding schemes and feature selection

We used eight encoding schemes to select features in the protein sequences. The schemes were composition of amino acids (AAs), composition of K-spaced amino acid pairs (CKSAAP), composition of motif or domain (PROSITE), pseudo amino acid composition (PseAAC) and combined methods, AA_CKSAAP, AA_PROSITE, CKSAAP_PROSITE and CKSAAP_AA_PROSITE.

Composition of AA

The frequency of one AA in sequence fragment was calculated by the following equation:

$$vi = \frac{ci}{len(seq)}, i = 1, \ldots, 20,$$

where Ci and len (*seq*) denote the composition of the corresponding AA in the sequence fragment and the length of the sequence fragment, respectively. v_i illustrates the frequency of the AAs in the protein sequence.

Composition of K-spaced amino acid pairs

CKSAAP has been successfully used to represent the sequence fragment [14, 15]. A sequence fragment may contain 400 types (AxA, AxC, AxD, ..., OxO) of K-spaced AA pairs (i.e. the pairs separated by K other AAs). The value of N_{AA} is the composition of the corresponding AA pairs in the sequence fragment, whereas N_{total} represent the total composition of AA pairs in the sequence fragment. The flowchart and the calculation used for the CKSAAP feature selection approach are shown in Lin *et al.* [16].

When the value of K increases, the prediction accuracy and the sensitivity increase, but the computational complexity and the required time for training the models also increase [14]. In this article, we considered the CKSAAP encoding scheme with k = 0, 1, 2, 3, 4 and 5, and the total dimension of the six-spaced feature vector is 2400.

Composition of motif or domain

We used perl script, ps_scan (ftp://ftp.expasy.org/databases/prosite/ps_scan/), to search the motif or domain in the sequence fragment in the PROSITE database, and then we calculated the frequency of the corresponding motif in the sequence fragment as the following:

$$vi = \frac{ci}{N_{entries}}, i = 1, \ldots, 2342,$$

where Ci denotes the composition of the corresponding motif or domain in the protein sequence fragment. N_{entries} denotes the number of all the motif or domain in the PROSITE database (total 2342 entries in prosite.dat Ver 20.83). v_i illustrates the frequency of the corresponding motif or domain in the protein sequence. The total dimension of PROSITE is 2400.

PseAAC

PseAAC was improved by Chou in 2005 and could be used to represent sequence-order or position-specific information of one protein or peptide [17, 18]. PseAAC for a protein or peptide P can be generally formulated as follows:

$$\mathrm{P} = [\psi_1 \psi_2 \psi_3 \ldots \psi_{\mathrm{u}} \ldots \psi_\Omega]^{\mathrm{T}}$$

where T is the transpose operator, whereas Ω is an integer to reflect the vector's dimension. In this research work, PseAAC-builder was downloaded and run to generate PseAAC information from the data set [19], whereas lamda parameter was set from 1 to 50 to get the optimal performance.

Combined methods

AA, CKSAAP and PROSITE were used to compose combined feature selection methods. Because of the high dimensionality of the CKSAAP and PROSITE encoding schemes, Relief-F was used to decrease the total dimension of combined methods. Each feature input was ranked and weighted using the K-nearest neighbors (KNNs) classification, and the features with positive weight were selected for the data set. The total dimension of the combination of CKSAAP_PROSITE was 1560, whereas that of AA_CKSAAP_PROSITE was 1573.

Classifiers

We compared four classification methods: naive Bayes (NB); KNN; RF; and support vector machine (SVM). These classifiers were implemented using the Waikato Environment for Knowledge Analysis software [20].

Naive Bayes

NB assumes the predictors are statistically independent, which makes it a classification tool that is easy to interpret. Because the inputs are assumed to be independent given the class, the conditional probability is calculated by using Bayes' theorem:

$$p(\mathrm{C}|F_1, F_2, \ldots, F_n) = \frac{p(\mathrm{C}) \prod_{\mathrm{i}=1}^{n} p(F_{\mathrm{i}}|\mathrm{C})}{p(F_1, F_2, \ldots, F_n)}$$

where *F* denotes the random variable corresponding to the input of the classifier and *C* denotes the binary random variable corresponding to the output of the classifier.

K-nearest neighbor

KNN rule is one of the simplest but powerful methods for performing nonparametric classification [21]. The KNN classifier has been successfully used to predict protein function [22], protein subcellular localization [23] and membrane protein type [24].

KNN classifies a new instance by evaluating its distance from each of the classifier instances and chooses the class label of the classifier instance that is closest to the new instance as the predicted class of the new instance. In this article, the distance (D) was calculated as following:

$$D = \sqrt{(x_1^{(1)} - x_1^{(2)})^2 + (x_2^{(1)} - x_2^{(2)})^2 + \ldots + (x_n^{(1)} - x_n^{(2)})^2}$$

where $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$ is the feature of a new instance, and $x_1^{(2)}$, $x_2^{(2)}, \ldots, x_n^{(2)}$ is the feature of another instance (Supplementary Figure S1).

Random forest

RF is an ensemble of unpruned decision trees [25], and has already been used to predict protein–protein interaction [26] and protein long disordered region [27]. In RF, the number of trees in the forest is adjustable, and each tree is grown to full length using a subset of the training data set. To classify an instance of unknown class label, each tree casts a unit classification vote. The forest selects the classification having the most votes over all the trees in the forest. Therefore, there are two key parameters in RF. One is the number of the trees, M, and the other is the number of features selected randomly, m. In this article, we selected the optimal value of M = 100, and determined m based on the result of a preliminary evaluation (Supplementary Figure S2).

Support vector machine

SVM is a popular machine learning algorithm mainly used to deal with binary classification problems. In this article, LibSVM under Weka with radial basis kernels was used as K $(x_i, y_i) = \exp(-\gamma \|x_i - y_i\|^2)$ [14]. We used grid search strategy to find the optimal parameters C ϵ $\{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \, \epsilon$ $\{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{3}\}$, and the total number of grids was 11*10 = 110. After training with the subset of the training data, the accuracy (ACC) of SVM predictor of every grid was calculated and compared (Supplementary Figure S3) to optimize the C and γ for SVM.

Performance measurement

Four measurements—sensitivity (*Sn*), specificity (*Sp*), accuracy (*ACC*) and the Matthew's correlation coefficient (*MCC*)—were used to evaluate the performance of the different predictors [28], which were defined below.

$$Sn = \frac{TP}{TP + FN},$$

$$Sp = \frac{TN}{TN + FP},$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

and

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

where *TP*, *FP*, *FN* and *TN* denote true positives, false positives, false negatives and true negatives.

We used SPSS 16.0 to create receiver operating characteristic (ROC) curves to compare the performance of different predictors. For each possible threshold, the sensitivity and specificity were evaluated, the ROC curve [sensitivity versus (1-specificity) curve] was plotted and the area underneath this curve was used to compare the performance of predictors with different feature selection methods.

Results and discussion

Evaluation on the performance of different encoding schemes for feature selection

Four different classifiers corresponding to eight feature selection methods were trained and used to predict G protein effectors. We first evaluated the predicting performance of different encoding schemes. The results are shown in Table 1 and Figure 1.

Among all the NB predictors, the NB with AA for feature selection achieved the highest ACC of 66.78%. The best prediction

Table 1. Predicting performance of NB, KNN, RF and SVM on the effectors of heterotrimeric G proteins in rice with different features selection methods

Method	Feature	Sp (%)	Sn (%)	ACC (%)	MCC
NB	AA	49.52±4.17	84.04±2.44	66.78±1.74	0.36±0.033
KNN (K = 7)		59.23±4.62	76.35±2.73	67.79±2.80	0.36±0.055
RF (m = 3)		73.75±3.10	68.75±2.07	71.25±2.13	0.43±0.043
SVM (c = 3, γ = 3)		72.60±2.83	72.02±4.02	72.31±2.94	0.45±0.059
NB	CKSAAP	40.38±4.15	87.02±0.89	63.70±2.25	0.31±0.045
KNN (K = 49)		64.33±8.58	70.77±6.79	67.55±3.76	0.35±0.075
RF (m = 50)		68.17±3.56	72.88±2.75	70.53±2.31	0.41±0.046
SVM (c = 1, γ = 3)		70.77±3.58	73.56±3.23	72.16±3.04	0.44±0.061
NB	PROSITE	31.83±13.76	80.87±7.35	56.35±3.55	0.14±0.067
KNN (K = 27)		74.13±7.54	48.75±5.00	61.44±2.90	0.23±0.064
RF (m = 56)		64.42±2.62	59.33±3.68	61.88±1.92	0.24±0.038
SVM (c = 15, γ = 3)		59.62±3.94	67.02±2.85	63.32±1.22	0.27±0.024
NB	PseAAC	48.46±0.08	75.96±0.05	62.21±0.03	0.25±0.056
KNN (K = 56)		52.40±0.09	72.60±0.05	62.50±0.04	0.25±0.072
RF (m = 13)		61.54±0.03	68.17±0.03	64.86±0.03	0.29±0.050
SVM (c = 0.5, γ = 0.0004)		63.46±0.04	64.42±0.03	63.94±0.03	027±0.058
NB	AA _CKSAAP	40.58±4.14	87.02±0.89	63.80±2.26	0.31±0.045
KNN (K = 44)		63.85±8.43	70.96±6.60	67.40±3.49	0.35±0.070
RF (m = 43)		68.17±2.33	74.42±2.51	71.30±2.30	0.43±0.046
SVM (c = −1, γ = 3)		72.98±3.71	72.88±3.35	72.93±2.90	0.46±0.058
Naive Bayes	AA_PROSITE	45.29±6.01	84.04±2.99	64.66±2.43	0.32±0.045
KNN (K = 17)		64.71±7.37	66.25±6.94	65.48±4.02	0.31±0.081
RF (m = 30)		76.06±3.17	68.17±2.41	72.12±1.86	0.44±0.038
SVM (c = 15, γ = -15)		72.12±3.49	71.83±2.89	71.97±2.92	0.44±0.058
NB	CKSAAP_PROSITE_1560	41.54±3.62	86.15±1.07	63.85±1.86	0.31±0.036
KNN (K = 37)		69.62±9.57	71.25±10.04	70.43±3.14	0.42±0.062
RF (m = 59)		71.25±3.74	73.85±2.42	72.55±1.89	0.45±0.037
SVM (c = 1, γ = 3)		73.17±4.33	73.37±3.16	73.27±3.06	0.47±0.061
NB	CKSAAP_PROSITE_200	44.13±4.33	87.69±0.26	65.91±1.40	0.35±0.025
KNN (K = 63)		74.90±6.28	69.04±4.01	71.97±1.97	0.44±0.042
RF (m = 11)		73.94±2.41	75.29±2.43	74.62±2.01	0.49±0.040
SVM (c = 8, γ = 8)		74.81±1.96	71.44±2.02	73.13±1.64	0.46±0.033
NB	AA_CKSAAP_PROSITE_1573	40.87±3.58	85.67±0.91	63.27±1.90	0.29±0.037
KNN (K = 33)		70.96±8.16	68.17±10.73	69.57±2.94	0.40±0.055
RF (m = 50)		71.44±4.79	73.37±2.58	72.40±2.65	0.45±0.053
SVM (c = -1, γ = 3)		73.46±4.15	73.56±2.76	73.51±3.00	0.47±0.060

Sp: Specificity; Sn: Sensitivity; ACC: Accuracy; MCC: Matthew's Correlation Coefficient. AA: Composition of amino acid; CKAAP: Composition of K-Spaced Amino Acid Pairs; PROSITE: Composition of motif or domain. AA_CKSAAP: Combined CKSAAP and AA; AA_PROSITE: Combined AA and PROSITE; CKSAAP_PROSITE_1560: Combined CKSAAP and PROSITE with 1560 dimensionality. AA_CKSAAP_ PROSITE_1573: Combined AA, CKSAAP and PROSITE with 1573 dimensionality. The same applies for Tables 2 and 3.

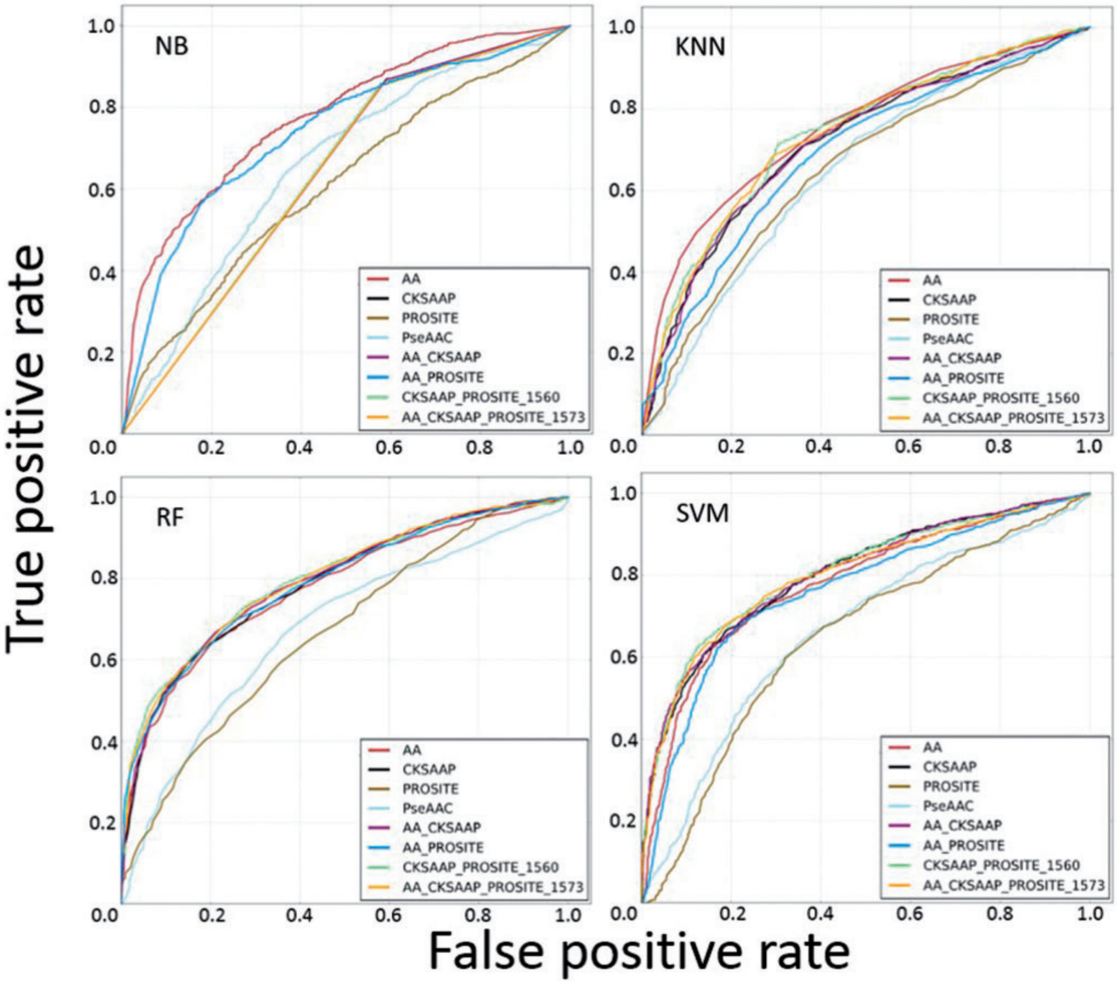

Figure 1. ROC curves of different NB, KNN, RF and SVM predictors on the effectors of heterotrimeric G proteins in rice with different features extracting methods.

performance of KNN predictors was obtained with CKSAAP_ PROSITE_1560 for feature selection. Within the different encoding schemes, the RF predictor achieved the best performance when CKSAAP_PROSITE_1560 was used for feature selection, whereas a SVM combined with AA_CKSAAP_ PROSITE _1573 for feature selection reached the best prediction performance (Table 1 and Figure 1). PseAAC did not outperform CKSAAP (Table 1), and each of them can represent sequence-order or position-specific information. Therefore, in this research work, we used AA, CKSAAP and PROSITE to compose the combined feature selection methods.

These results indicate that the different encoding schemes for feature selection in KNN, RF or SVM predictors were complementary to some extent. This is owing to the different ability of different sole encoding schemes in extracting the character of protein sequences. The AA encoding scheme clearly characterizes AAs in different positions of the protein sequences, CKSAAP reflects the relationship between AA pairs at different positions [16, 29] and PROSITE illustrates the frequency of the corresponding motif or domain. In our previous research, we found that AA and CKSAAP showed complementary capability in extracting the sequence character surrounding a potential phosphorylated site [7]. On the other hand, protein–protein interactions are frequently mediated by the binding of a modular domain in one protein to a short, linear motif in its partner [30]. The AA sequence of a domain and the characteristics of its ligand-binding site determined the intrinsic specificity of a modular domain, which are context-independent because they are retained even in the isolated domains [31]. It could be hypothesized that the AAs surrounding or inside of a modular domain or motif contribute to protein binding specificity. Therefore, CKSAAP and PROSITE complement each other in extracting the sequence character of a modular domain or motif, which might be related to the specificity of effectors binding to G protein. We highlight here that the combination of sequence and domain features contributes to the final improvement on predicting the partners of one protein.

Evaluation on the performance of different classifiers

The ACC and MCC of RF and SVM predictors were higher than that of NB and KNN predictors with different feature selection methods (Figure 2). RF reached its highest ACC when CKSAAP_PROSITE_1560 was used for feature selection method, whereas SVM achieved its highest ACC when AA_CKSAAP_ PROSITE_ 1573 was used for feature selection (Figure 2). MCC value derived by Jackknife test shows the same trend (Figure 2).

It is fair to compare the classifiers not only by the average ACC of prediction but also by the trade-off between Sn and Sp [32]. The gap between the sensitivity (Sn) and the specificity (Sp) of RF and SVM predictors was always lower than that of NB and

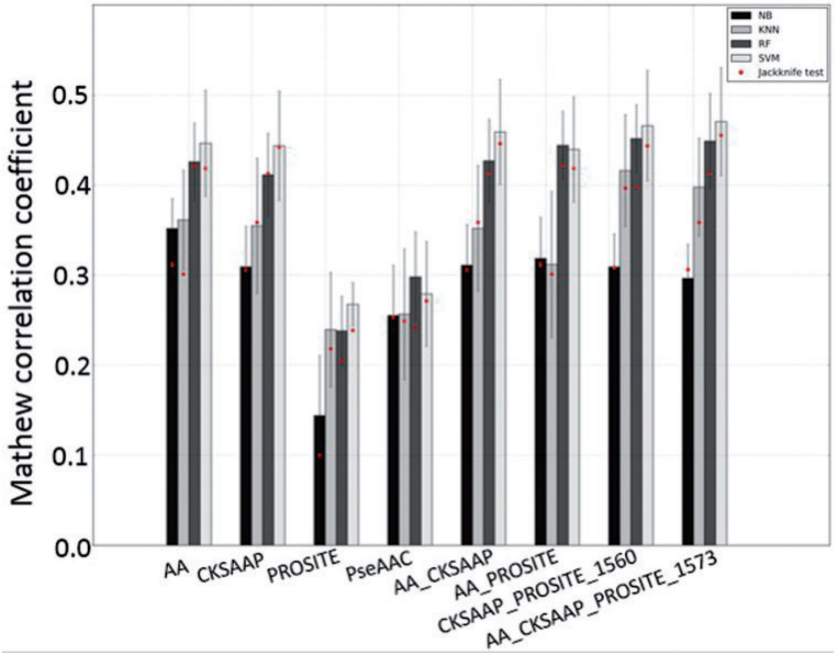

Figure 2. Predicting performance of different classification methods with different feature extraction methods. The performance is examined by using 10-fold cross-validation and jackknife test. The error bars indicate the standard deviations of MCC values for the case of 10-fold cross-validation with 10 runs.

KNN. This result implied that RF and SVM classifier performed better on the G protein effectors than NB and KNN classifiers. This is owing to the different ability of these four common-used classifiers in dealing with multidimensional data sets. NB classifier has strong independent assumption of the feature variables so that only a small training sample size is needed to represent the feature space [33]. In reality, the multidimensional data sets are seldom independent. The low ACC of KNN might be a result of inadequate size of the training data set. However, RF and SVM classifiers can deal with data set suffering heavily from high-dimensional, noisy, with missing values, categorical and highly correlated features [34].

RF classifier combined with CKSAAP_PROSITE_200 for feature selection yielded the best performance

RF variable importance measures rankings can be used for screening or filtering by selecting top-ranking parameters for follow-up study [35]. The RF classifier was adopted as the prediction engine and operated with the optimal feature selection method after taking both the prediction performance and the capability of ranking features into consideration.

Relief is a feature weight-based algorithm that can detect those features that are statistically relevant to the target concept [36]. Relief-F is the extension to the original Relief algorithm and is able to deal with noisy and multi-class problems rather than two-class problem [37]. We applied Relief-F to determine the optimal features number for RFs. The individual RF predictors corresponding to different feature subsets were constructed and examined by using the 10-fold cross-validation on the benchmark data set and setting 1500 for sample size (m) and 100 for the feature-increasing gap. Figure 3 showed that MCC of the corresponding predictor declined rapidly as dimensionality increased. This was consistent with the research of Winham *et al.* [38], which reported that the ability of RF to detect SNP effects diminished as dimensionality increased.

RF combined with CKSAAP_PROSITE for feature selection achieved the highest MCC, 0.49 when 200 features were included (Table 1 and Figure 3). The Sn, Sp and ACC were 75.29%, 73.94% and 74.62%, respectively. Therefore, in this study, we used RF classifier combined with CKSAAP_PROSITE_200 for feature selection to develop a G-effector predictor. Our predictor is accessible at http://bioinformatics.fafu.edu.cn/geffector.

G-effector outperforms BLAST, PSI-BLAST and HMMER

A comparison between the results of our G-Effector with BLAST, PSI-BLAST and HMMER predictors were examined using a 10-fold cross-validation data set. First, the data were divided equally into 10-folds, 1-fold of data was used as the test data and the remaining 9-folds of data as the training data. Second, we optimized the E-value, and ran PSI-BLAST with three-times iteration. The result showed that ACC of G-Effector, BLAST, PSI-BLAST and HMMER were 74.62%, 59.81%, 66.54% and 57.55%, respectively (Table 2). G-Effector also outperforms the traditional similarity search tools on an independent test data (Figure 4).

Prediction of the G protein effectors in rice

We used G-Effector to predict the effectors of heterotrimeric G proteins in rice. First, rice proteome sequences were downloaded from Rice Genome Annotation Project (RGAP, http://rice.plantbiology.msu.edu/) [39]. All of the rice protein sequences were run through the G-Effector predictor, and the top 30 predicted effectors were selected for follow-up analysis. Interacting gene partners typically have similar expression profiles over many conditions [40], and so, we checked whether the candidate effectors co-expressed with RGA1, RGB1, RGG1 or RGG2 seen in transcriptomics data from ROAD (Rice Oligonucleotide Array Database). The top 15 candidate effectors strongly co-expressing with RGA1 or RGB1 are presented in Table 3. Interestingly, one of the top 15 candidate effectors, LOC_Os06g48590, had been reported to interact with Gß in rice under stress [12].

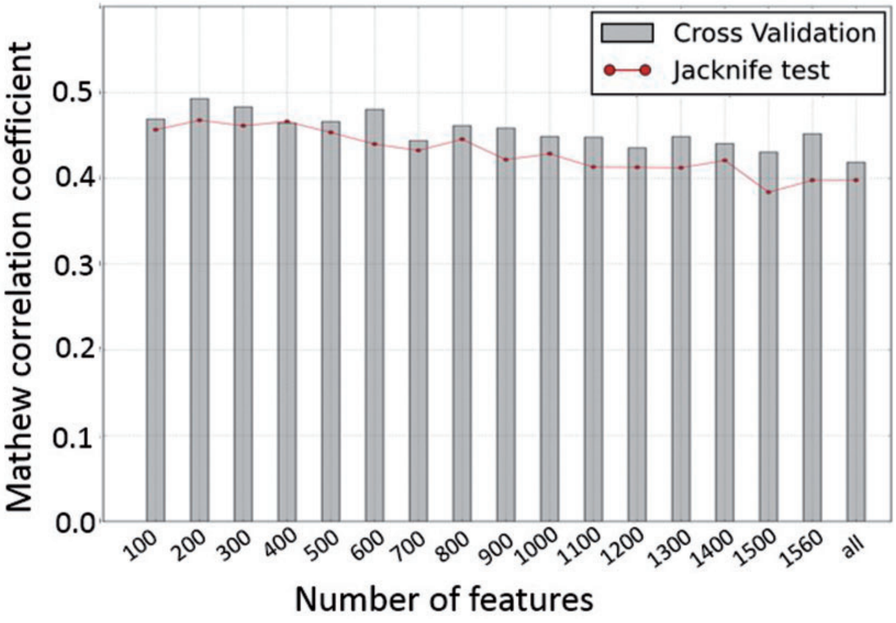

Figure 3. Feature optimization by using Relief-F selection method. Each performance is examined by jackknife test.

Table 2. Predicting performance of G-effector, BLAST, PSI-BLAST and HMMER, on the effectors of heterotrimeric G proteins in rice

Methods	Parameter	Sp (%)	Sn (%)	ACC (%)	MCC
G-Effector	m = 11	73.94±2.41	75.29±2.43	74.62±2.01	0.49±0.040
BLAST	e-value = 0.5	73.46±3.39	46.15±0.00	59.81±1.70	0.20±0.037
PSI-BLAST	e-value = 0.5	73.46±3.39	59.62±0.00	66.54±1.70	0.33±0.036
HMMER	e-value = 0.1	88.17±2.72	26.92±0.00	57.55±1.36	0.19±0.039

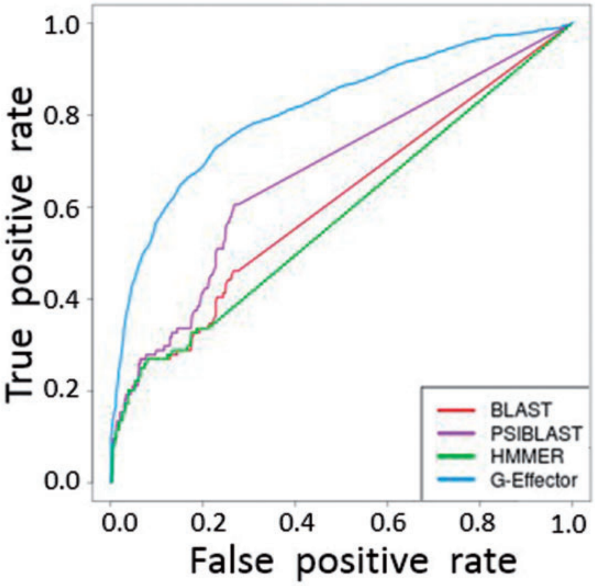

Figure 4. ROC curves of G-Effector, BLAST, PSI-BLAST and HMMER on the effectors of heterotrimeric G proteins in rice.

Table 3. Effectors of heterotrimeric G proteins in rice predicted by G-effectors and verified by gene co-expression data

No.	Subunit	Effectors	PCC	Score[a]
1	Gα	LOC_Os06g34690	0.51	0.89
2	Gα	LOC_Os04g46620	0.50	0.87
3	Gα	LOC_Os10g10244	0.50	0.85
4	Gα	LOC_Os01g19450	0.50	0.69
5	Gα	LOC_Os02g05630	0.59	0.59
6	Gβ	LOC_Os06g34690	0.80	0.89
7	Gβ	LOC_Os04g46620	0.72	0.87
8	Gβ	LOC_Os03g59020	0.69	0.87
9	Gβ	LOC_Os03g64210	0.57	0.85
10	Gβ	LOC_Os05g28280	0.66	0.85
11	Gβ	LOC_Os06g47320	0.78	0.83
12	Gβ	LOC_Os10g32550	0.54	0.83
13	Gβ	LOC_Os06g45710	0.55	0.81
14	Gβ	LOC_Os06g48590	0.64	0.58
15	Gβ	LOC_Os10g08550	0.68	0.79

[a]Score: predicted by G-Effector tool.

Conclusion

In this article, we first compared the performance of different classifier combined with different encoding schemes for feature selection by using 10-fold cross-validation data set. RF classifier was adopted as the prediction engine because of its predicted performance and its capability of ranking features. Relief-F was applied to determine the optimal feature number for RF classifier. RF combined with CKSAAP_PROSITE for feature selection achieved a maximum of MCC equaling 0.49 when 200 features were included.

The Web server, G-Effector, was developed using RF classifier combined with CKSAAP_PROSITE_200 for feature selection and is freely accessible at http://bioinformatics.fafu.edu.cn/gef fector. The G-Effector predictor outperformed the existing three similarity search tools when tested by an independent data set.

We used G-Effector to screen the effectors of heterotrimeric G proteins in rice, and we confirmed the candidate effectors by using gene co-expression data. One of the top 15 candidate effectors is verified by the research work of Bhardwaj *et al.* [12]. The candidate effectors' list in this article provides both a clue for researchers as to their function and a framework of validation for future experimental work.

Key Points

- There are biological, technical and experimental needs to evaluate the predicting performance of different classifiers combined with different feature selection methods in predicting the effectors of heterotrimeric G proteins and use the best one to develop a new online predictor.
- Compared with other algorithms, RF classifier combined with CKSAAP_PROSITE_200 for feature selection yields the best performance.
- An online predictor, G-Effector, is developed by using RF classifier combined with CKSAAP_PROSITE_200 for feature selection method.
- G-Effector outperforms the traditional similar search tools, including BLAST, PSI-BLAST and HMMER, on predicting G protein effectors.

Supplementary data

Supplementary data are available online at http://bib.oxford journals.org/.

Acknowledgment

The authors thank the anonymous referees whose constructive comments were helpful in improving the quality of this work.

Funding

This work was supported by Natural Science Foundation of China (No. 31270454 and 81502091), Natural Science Foundation of Fujian, China (No. 2013J01077), the Key Program of Ecology in Fujian, China (No. 0608507 and 0b08b005).

References

1. Yadav DK, Tuteja N. Rice G-protein coupled receptor (GPCR): *in silico* analysis and transcription regulation under abiotic stress. *Plant Signal Behav* 2011;**6**:1079–86.
2. Gookin TE, Kim J, Assmann SM. Whole proteome identification of plant candidate G-protein coupled receptors in Arabidopsis, rice, and poplar: computational prediction and *in-vivo* protein coupling. *Genome Biol* 2008;**9**:R120.
3. Theodoropoulou MC, Bagos PG, Spyropoulos IC, *et al.* gpDB: a database of GPCRs, G-proteins, effectors and their interactions. *Bioinformatics* 2008;**24**:1471–2.
4. McCudden CR, Hains MD, Kimple RJ, *et al.* G-protein signaling: back to the future. *Cell Mol Life Sci* 2005;**62**:551–77.
5. Jones AM, Assmann SM. Plants: the latest model system for G-protein research. *EMBO Rep* 2004;**5**:572–8.
6. Apone F, Alyeshmerni N, Wiens K, *et al.* The G-protein-coupled receptor GCR1 regulates DNA synthesis through activation of phosphatidylinositol-specific phospholipase C. *Plant Physiol* 2003;**133**:571–9.
7. Liu X, Yue Y, Li B, *et al.* A G protein-coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid. *Science* 2007;**315**:1712–6.
8. Gao Y, Zeng Q, Guo J, *et al.* Genetic characterization reveals no role for the reported ABA receptor, GCR2, in ABA control of seed germination and early seedling development in Arabidopsis. *Plant J* 2007;**52**:1001–13.
9. Fredriksson R, Schioth HB. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* 2005;**67**:1414–25.
10. Moriyama EN, Strope PK, Opiyo SO, *et al.* Mining the Arabidopsis thaliana genome for highly-divergent seven transmembrane receptors. *Genome Biol* 2006;**7**:R96.
11. Friedman EJ, Wang HX, Jiang K, *et al.* Acireductone dioxygenase 1 (ARD1) is an effector of the heterotrimeric G protein beta subunit in Arabidopsis. *J Biol Chem* 2011;**286**:30107–18.
12. Bhardwaj D, Sheikh AH, Sinha AK, *et al.* Stress induced beta subunit of heterotrimeric G-proteins from Pisum sativum interacts with mitogen activated protein kinase. *Plant Signal Behav* 2011;**6**:287–92.
13. Jia J, Liu Z, Xiao X, *et al.* iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules* 2016; **21**:95.
14. Zhao X, Zhang W, Xu X, *et al.* Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs. *PLoS One* 2012;**7**:e46302.
15. Chen YZ, Tang YR, Sheng ZY, *et al.* Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics* 2008;**9**:101.
16. Lin S, Song Q, Tao H, *et al.* Rice_Phospho 1.0: a new rice-specific SVM predictor for protein phosphorylation sites. *Sci Rep* 2015;**5**:11940.
17. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;**21**:10–9.
18. Chou K-C. Impacts of bioinformatics to medicinal chemistry. *Med Chem* 2015;**11**:218–34.
19. Du P, Wang X, Xu C, *et al.* PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal Biochem* 2012;**425**: 117–9.
20. Witten IH, Frank E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, CA, 2005.
21. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* 2011;**273**:236–47.
22. Lan L, Djuric N, Guo Y, *et al.* MS-kNN: protein function prediction by integrating multiple data sources. *BMC Bioinformatics* 2013;**14 Suppl 3**:S8.
23. Chou KC, Shen HB. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. *J Proteome Res* 2006;**5**:1888–97.
24. Shen H, Chou KC. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid

composition to predict membrane protein types. *Biochem Biophys Res Commun* 2005;**334**:288–92.

25. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
26. Li BQ, Feng KY, Chen L, *et al*. Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. *PLoS One* 2012;**7**:e43927.
27. Han P, Zhang X, Norton RS, *et al*. Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinformatics* 2009;**10**:8.
28. Que S, Li K, Chen M, *et al*. PhosphoRice: a meta-predictor of rice-specific phosphorylation sites. *Plant Methods* 2012;**8**:1–9.
29. Chen Z, Chen YZ, Wang XF, *et al*. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One* 2011;**6**:e22930.
30. Li L, Zhao B, Du J, *et al*. DomPep–a general method for predicting modular domain-mediated protein-protein interactions. *PLoS One* 2011;**6**:e25528.
31. Pawson T, Nash P. Assembly of cell regulatory systems through protein interaction domains. *Science* 2003;**300**: 445–52.
32. Song K, Zhang Z, Tong TP, *et al*. Classifier assessment and feature selection for recognizing short coding sequences of human genes. *J Comput Biol* 2012;**19**:251–60.
33. Entezari-Maleki R, Rezaei A, Minaei-Bidgoli B. Comparison of classification methods based on the type of attributes and sample size. *J Convergence Inf Technol* 2009;**4**:94–102.
34. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:27.
35. Schwarz DF, Szymczak S, Ziegler A, *et al*. Picking single-nucleotide polymorphisms in forests. *BMC Proc* 2007;**1 Suppl 1**:S59.
36. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. *AAAI* 1992;129–34.
37. Chang SW, Abdul-Kareem S, Merican AF, *et al*. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics* 2013;**14**:170.
38. Winham SJ, Colby CL, Freimuth RR, *et al*. SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinformatics* 2012;**13**:164.
39. Ouyang S, Zhu W, Hamilton J, *et al*. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res* 2007;**35**:D883–7.
40. Fraser HB, Hirsh AE, Wall DP, *et al*. Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA* 2004;**101**:9033–8.